APPENDIX

A. Proof of Lemma 1

Proof. Let $P_e = \Pr(\hat{M} \neq M)$. From the Fano's inequality, we have

$$\mathsf{H}(M|\hat{M}, \zeta^T) \leq \mathsf{H}(M|\hat{M}) \leq 1 + P_e \log m.$$

The entropy of M is upper bounded by

$$\begin{aligned}\log m = \mathsf{H}(M) = \mathsf{H}(M|\zeta^T) &= \mathsf{I}(M; \hat{M}|\zeta^T) + \mathsf{H}(M|\hat{M}, \zeta^T) \\ &\leq \mathsf{I}(M; X^T|\zeta^T) + 1 + P_e \log m \\ &\leq H(X^T|\zeta^T) + 1 + P_e \log m,\end{aligned}$$

which leads to

$$\frac{\log m}{T} \leq \frac{H(X^T|\zeta^T)}{T} + \frac{1}{T} + P_e \frac{\log m}{T}.$$

If $P_e \to 0$ as $T \to \infty$, we have

$$\frac{\log m}{T} \leq \frac{H(X^T|\zeta^T)}{T} \leq \mathsf{H}(P_X) \leq \sup_{P_X : \mathsf{D}(P_X^T, Q_X^T) \leq d} \mathsf{H}(P_X).$$

□

B. Proof of Lemma 2

Proof. For any $i \neq j$, define the relative entropy typical set

$$\mathcal{A}_{\epsilon,i,j}^{(T)}(\mathbb{P}_i \| \mathbb{P}_j) \coloneqq \left\{ (x^T, \zeta^T) : \left| \frac{1}{T} \log \frac{\mathbb{P}_i(x^T, \zeta^T)}{\mathbb{P}_j(x^T, \zeta^T)} - \mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) \right| \leq \epsilon \right\}.$$

We have $\mathbb{P}_j(\mathcal{B}_{T,j}^{\mathrm{c}}) = 1 - \mathbb{P}_j(\mathcal{B}_{T,j})$ and

$$\begin{aligned}\mathbb{P}_j(\mathcal{B}_{T,j}) &= 1 - \sum_{i:i\neq j} \mathbb{P}_j(\mathcal{B}_{T,i}) \leq 1 - \sum_{i:i\neq j} \mathbb{P}_j(\mathcal{B}_{T,i} \cap \mathcal{A}_{\epsilon,i,j}^{(T)}) \\ &\leq 1 - \sum_{i:i\neq j} \sum_{(x^T,\zeta^T) \in \mathcal{B}_{T,i} \cap \mathcal{A}_{\epsilon,i,j}^{(T)}} \mathbb{P}_i(x^T, \zeta^T) \exp(-T(\mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon)) \\ &= 1 - \sum_{i:i\neq j} \exp(-T(\mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon)) \mathbb{P}_i(\mathcal{B}_{T,i} \cap \mathcal{A}_{\epsilon,i,j}^{(T)}) \\ &\overset{\text{(a)}}{\leq} 1 - \sum_{i:i\neq j} \exp(-T(\mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon))(1 - 2\epsilon) \\ &\leq 1 - m(1-2\epsilon) \exp(-T(\min_{i:i\neq j} \mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon)) \\ &\leq 1 - m(1-2\epsilon) \exp(-T(\max_{P_X : \mathsf{D}(P_X^T, Q_X^T) \leq d} \min_{i:i\neq j} \mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon))\end{aligned}$$

where (a) follows since $\mathbb{P}_i(\mathcal{B}_{T,i} \cap \mathcal{A}_{\epsilon,i,j}^{(T)}) = 1 - \mathbb{P}_i(\mathcal{B}_{T,i}^{\mathrm{c}} \cup (\mathcal{A}_{\epsilon,i,j}^{(T)})^{\mathrm{c}}) \geq 1 - \mathbb{P}_i(\mathcal{B}_{T,i}^{\mathrm{c}}) - \mathbb{P}_i((\mathcal{A}_{\epsilon,i,j}^{(T)})^{\mathrm{c}}) \geq 1 - 2\epsilon$ for sufficiently large T. The proof is thus complete. □

C. Proof of Theorem 3

a) Existence of asymptotically optimal decoders: First, the function g proposed in Theorem 3 always exists, as discussed in Remark 1. If the number of message bits satisfies $\frac{1}{T}(\log m - \log \alpha) \leq \mathsf{H}(P_X^*)$, then we have

$$m \leq e^{T\mathsf{H}(P_X^*)} \doteq \mathcal{A}_{\eta,X}^{(T)},$$

and the output space of g contains $[m]$. Thus, any decoder in the class of asymptotically optimal decoders Γ_η^* can decode messages drawn from $[m]$.

b) Asymptotic optimality: For any $\gamma \in \Gamma_\eta^*$, one can always construct the corresponding encoder outputs $P_{X,\zeta|M}^*$ in Theorem 3. In the following, we first show that the probability of the atypical set decays exponentially with T. We then prove that the j-th error probability vanishes to 0 while the worst-case false alarm error is upper bounded by α as $T \to \infty$.

Let $\eta = T^{-\frac{1}{4}}$ and define the set $\mathcal{A}_{\eta,j}^{(T)}$ of jointly typical sequences $\{(x^T, \zeta^T)\}$ w.r.t. the distribution $P_{X,\zeta|M=j}$ as

$$\begin{aligned}\mathcal{A}_{\eta,j}^{(T)} \coloneqq \Bigg\{ (x^T, \zeta^T) \in \mathcal{X}^T \times \mathcal{Z}^T : &\left| -\frac{1}{T} \log P_X^T(x^T) - \mathsf{H}(P_X) \right| \leq \eta, \left| -\frac{1}{T} \log P_\zeta^T(\zeta^T) - \mathsf{H}(P_\zeta) \right| \leq \eta, \\ &\left| -\frac{1}{T} \log P_{X,\zeta|M=j}^T(x^T, \zeta^T) - \mathsf{H}(P_{X,\zeta|M=j}) \right| \leq \eta \Bigg\}.\end{aligned}$$

First, we bound the probability of the atypical sets $(\mathcal{A}_{\eta,X}^{(T)})^{\mathrm{c}}, (\mathcal{A}_{\eta,\zeta}^{(T)})^{\mathrm{c}}, (\mathcal{A}_{\eta,j}^{(T)})^{\mathrm{c}}$. From the union bound, we have

$$\begin{aligned}\mathbb{P}_j((X^T,\zeta^T)\notin\mathcal{A}_{\eta,j}^{(T)}) \le \mathbb{P}_j\bigg(\bigg|-\frac{1}{T}\log P_X^T(x^T)-\mathsf{H}(P_X)\bigg|\ge\eta\bigg)+\mathbb{P}_j\bigg(\bigg|-\frac{1}{T}\log P_\zeta^T(\zeta^T)-\mathsf{H}(P_\zeta)\bigg|\ge\eta\bigg)\\ +\mathbb{P}_j\bigg(\bigg|-\frac{1}{T}\log P_{X,\zeta|M=j}^T(x^T,\zeta^T)-\mathsf{H}(P_{X,\zeta|M=j})\bigg|\ge\eta\bigg). \qquad (3)\end{aligned}$$

Then, by the Chernoff bound, we have

$$\begin{aligned}\mathbb{P}_j\bigg(\bigg|-\frac{1}{T}\log P_X^T(x^T)-\mathsf{H}(P_X)\bigg|\ge\eta\bigg) &\le 2\mathbb{P}_j\bigg(-\frac{1}{T}\log P_X^T(x^T)-\mathsf{H}(P_X)\ge\eta\bigg)\\ &\le 2\exp\bigg(-T\sup_{s\ge0}(s\eta-\log\mathbb{E}[\exp(-s\log P_{X^T}(X^T))])\bigg)\\ &\overset{\text{(a)}}{\approx} 2\exp\bigg(-T\sup_{s\ge0}(s\eta-\big(-s\mathbb{E}[\log P_{X^T}(X^T)]+s^2\mathbb{E}[(\log P_{X^T}(X^T))^2]\big)\bigg)\\ &\overset{(b)}{=} 2\exp(-\Omega(T\eta^2))=\exp(-\Omega(T^{\frac{1}{2}})),\end{aligned}$$

where (a) follows from the Taylor expansion of $\exp(\cdot)$ and $\log(\cdot)$ and (b) follows since the maximum is achieved by $s=O(\eta)$. The rest of the terms in the union bound (3) can be similarly proved.

Thus, the probability of the jointly atypical set is upper bounded by

$$\mathbb{P}_j((X^T,\zeta^T)\notin\mathcal{A}_{\eta,j}^{(T)})\le 3\exp(-\Omega(T^{\frac{1}{2}}))=\exp(-\Omega(T^{\frac{1}{2}})).$$

Next, we prove that the proposed watermarking scheme in Theorem 3 achieves the asymptotic optimality. Let $P_X^*=Q_X$, $\mathcal{Z}\subset\mathbb{Z}$ and design $P_\zeta^*\in\mathcal{P}(\mathcal{Z})$ such that $\mathsf{H}(P_\zeta^*)=\mathsf{H}(P_X^*)$.

For any $\gamma^*\in\Gamma^*$, under the watermarking scheme given in Theorem 3, for any $j\in[m]$, the j-th error probability is given by

$$\begin{aligned}\beta_j(\gamma^*,P_{X^T,\zeta^T|M=j}^*) &= \sum_{x^T,\zeta^T}P_{X^T,\zeta^T|M}^*(x^T,\zeta^T|j)\mathbb{1}\{\gamma^*(x^T,\zeta^T)\ne j\}\\ &\le \sum_{(x^T,\zeta^T)\in\mathcal{A}_{\eta,j}^{(T)}}P_{X^T,\zeta^T|M}^*(x^T,\zeta^T|j)\mathbb{1}\{\gamma^*(x^T,\zeta^T)\ne j\}+\exp(-\Omega(T^{\frac{1}{2}}))\\ &= \exp(-\Omega(T^{\frac{1}{2}}))\to 0 \text{ as } T\to\infty.\end{aligned}$$

For $j=0$, the worst-case false alarm error probability is upper bounded as follows. For any $x^T\in\mathcal{A}_{\eta,X}^{(T)}$,

$$\begin{aligned}\sum_{\zeta^T}P_\zeta^*(\zeta^T)\mathbb{1}\{\gamma^*(x^T,\zeta^T)\ne0\} &\le \sum_{\zeta^T\in\mathcal{A}_{n,\zeta}^{(T)}}P_\zeta^*(\zeta^T)\mathbb{1}\{\gamma^*(x^T,\zeta^T)\ne0\}+\exp(-\Omega(T^{\frac{1}{2}}))\\ &= \sum_{i\in[m]}\sum_{\zeta^T\in\mathcal{A}_{n,\zeta}^{(T)}}P_\zeta^*(\zeta^T)\mathbb{1}\{\gamma^*(x^T,\zeta^T)=i\}+\exp(-\Omega(T^{\frac{1}{2}}))\\ &= \sum_{i\in[m]}\sum_{\zeta^T\in\mathcal{A}_{n,\zeta}^{(T)}}\bigg(\frac{1}{m}\sum_{j\in[m]}\sum_{x^T}P_{X^T,\zeta^T|M}^*(x^T,\zeta^T|j)\bigg)\mathbb{1}\{\gamma^*(x^T,\zeta^T)=i\}+\exp(-\Omega(T^{\frac{1}{2}}))\\ &\doteq \sum_{i\in[m]}\sum_{\zeta^T\in\mathcal{A}_{n,\zeta}^{(T)}}e^{-T\mathsf{H}(\zeta)}\mathbb{1}\{\gamma^*(x^T,\zeta^T)=i\}+\exp(-\Omega(T^{\frac{1}{2}}))\\ &= me^{-T\mathsf{H}(\zeta)}+\exp(-\Omega(T^{\frac{1}{2}}))\\ &\overset{\text{(a)}}{\le}\alpha+\exp(-\Omega(T^{\frac{1}{2}}))\\ &\xrightarrow{T\to\infty}\alpha,\end{aligned}$$

where (a) follows from the condition $\log m\le\log\alpha+T\mathrm{H}(P_\zeta^*)$ in Theorem 3.

For any $x^T\in(\mathcal{A}_{\eta,X}^{(T)})^{\mathrm{c}}$,

$$\sum_{\zeta^T}P_\zeta^*(\zeta^T)\mathbb{1}\{\gamma^*(x^T,\zeta^T)\ne0\}=0.$$

Since any distribution Q_X^T can be written as a linear combinations of $\{\delta_{x^T}\}_{x^T\in\mathcal{X}^T}$, we have

$$\sup_{Q_X}\beta_0(\gamma^*,Q_X\otimes P_\zeta^*)=\sup_{Q_X}\sum_{x^T,\zeta^T}Q_X^T(x^T)P_\zeta^*(\zeta^T)\mathbb{1}\{\gamma^*(x^T,\zeta^T)\ne0\}\to\alpha,\ \text{as } T\to\infty.$$

D. Proof of Theorem 4 and Theorem 5

We restate the optimization problem (P1) as follows:

$$\begin{aligned}\min_{\gamma,\mathbb{P}_1,\ldots,\mathbb{P}_m}\ \max_{j\in[m]}\ & \beta_j(\gamma, P_{X^T,\zeta^T|M=j})\\ \text{s.t.}\quad \sup_{Q_{X^T}}\ & \beta_0(\gamma, Q_{X^T}\otimes P_{\zeta^T})\le\alpha,\\ & \mathsf{D}(P_{X^T},Q_{X^T})\le d.\end{aligned}$$

Assumption 1 implicitly imposes the constraint that all $\mathbb{P}_1,\ldots,\mathbb{P}_m$ should have the same marginal distributions projected on $\mathcal{X}^T$ and on $\mathcal{Z}^T$, i.e., P_{X^T} and P_{ζ^T}.

a) Converse:

Proof of lower bound. First, let us fix a decoder γ. From the worst-case false alarm constraint, we have

$$\begin{aligned}\alpha\ge\sup_{Q_{X^T}}\beta_0(\gamma,Q_{X^T}\otimes P_{\zeta^T})&\ge\sum_{\zeta^T}P_{\zeta^T}(\zeta^T)\mathbb{1}\{\gamma(x^T,\zeta^T)\neq 0\}\\ &=\sum_{i\in[m]}\sum_{\zeta^T}P_{\zeta^T}(\zeta^T)\mathbb{1}\{\gamma(x^T,\zeta^T)=i\},\quad\forall x^T.\end{aligned}$$

Therefore, we have

$$\sum_{\zeta^T}P_{\zeta^T}(\zeta^T)\mathbb{1}\{\gamma(x^T,\zeta^T)=i\}\le\alpha_i,\quad\forall i,x^T,\quad\sum_{i\in[m]}\alpha_i=\alpha.\tag{4}$$

The j-th error probability is lower bounded by

$$\begin{aligned}\beta_j(\gamma,P_{X^T,\zeta^T|M=j})&=1-\mathbb{P}_j(\gamma(X^T,\zeta^T)=j)\\ &=1-\sum_{x^T,\zeta^T}P_{\zeta^T}(\zeta^T)P_{X^T|\zeta^T,M}(x^T|\zeta^T,j)\mathbb{1}\{\gamma(x^T,\zeta^T)=j\}.\end{aligned}$$

From (4), we have

$$\sum_{\zeta^T}P_{\zeta^T}(\zeta^T)P_{X^T|\zeta^T,M}(x^T|\zeta^T,j)\mathbb{1}\{\gamma(x^T,\zeta^T)=j\}\le\sum_{\zeta^T}P_{\zeta^T}(\zeta^T)\mathbb{1}\{\gamma(x^T,\zeta^T)=j\}\le\alpha_j,$$

and since $\mathbb{1}\{\gamma(x^T,\zeta^T)=j\}\le 1$,

$$\sum_{\zeta^T}P_{\zeta^T}(\zeta^T)P_{X^T|\zeta^T,M}(x^T|\zeta^T,j)\mathbb{1}\{\gamma(x^T,\zeta^T)=j\}\le\sum_{\zeta^T}P_{\zeta^T}(\zeta^T)P_{X^T|\zeta^T,M}(x^T|\zeta^T,j)=P_{X^T}(x^T).$$

Therefore, we lower bound β_j as follows

$$\begin{aligned}\beta_j(\gamma,P_{X^T,\zeta^T|M=j})&\ge 1-\sum_{x^T}(P_{X^T}(x^T)\wedge\alpha_j)=\sum_{x^T}(P_{X^T}(x^T)-\alpha_j)_+\\ &\ge\min_{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T})\le d}\sum_{x^T}(P_{X^T}(x^T)-\alpha_j)_+\quad\forall j\in[m].\end{aligned}$$

Among all possible $(\alpha_1,\ldots,\alpha_m)$ that sum up to α, the vector that minimizes the lower bound for $\max_{j\in[m]}\beta_j(\gamma,P_{X^T,\zeta^T|M=j})$ is $(\frac{\alpha}{m},\ldots,\frac{\alpha}{m})$. The proof is as follows:

$$\begin{aligned}\max_{j\in[m]}\beta_j(\gamma,P_{X^T,\zeta^T|M=j})&\ge\max_{j\in[m]}\min_{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T})}\sum_{x^T}(P_{X^T}(x^T)-\alpha_j)_+\\ &\overset{(a)}{\ge}\min_{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T})\le d}\sum_{x^T}\left(P_{X^T}(x^T)-\frac{\alpha}{m}\right)_+,\end{aligned}\tag{5}$$

where (a) holds with equality when $\alpha_j=\frac{\alpha}{m}$ for all $j\in[m]$.

We observe that the lower bound (5) is independent of γ. Thus, the lower bound also holds for the optimal value of the optimization problem (P1).

□

b) Achievability: Choose $\mathcal{Z}^T\subset\mathbb{Z}^T$ such that $|\mathcal{Z}|^T=|\mathcal{X}|^T+1$. Randomly pick a redundant sequence $\tilde{\zeta^T}\in\mathcal{Z}^T$. For any $m\le|\mathcal{X}|^T$, define a set of decoders as

$$\Gamma^*_{\tilde{\zeta^T}}\coloneqq\left\{\gamma\ \middle|\ \begin{array}{c}\gamma(x^T,\zeta^T)=\begin{cases}j,\quad\text{if }\zeta^T\neq\tilde{\zeta^T}\\ \quad\text{and } h(x^T,\zeta^T)=j\le m,,\\ 0,\quad\text{otherwise,}\end{cases}\\ \text{for some function } h:\mathcal{X}^T\times\mathcal{Z}^T\backslash\{\tilde{\zeta^T}\}\to[|\mathcal{X}^T|]\text{ satisfying that}\\ h(x^T,\cdot)\text{ and }h(\cdot,\zeta_1^T)\text{ are both bijective, given any fixed }x^T\text{ and fixed }\zeta_1^T\end{array}\right\}.$$

Construct $P^*_{\zeta^T|M} = P^*_{\zeta^T}$ as follows

$$P^*_{\zeta^T} = \left(\underbrace{\left(P^*_{X^T}(x^T) \wedge \frac{\alpha}{m}\right)_{x^T \in \mathcal{X}^T}}_{P^*_{\zeta^T}(\zeta^T),\ \forall \zeta^T \in \mathcal{Z}^T \setminus \{\tilde{\zeta^T}\}}, \underbrace{\sum_{x^T \in \mathcal{X}^T} \left(P^*_{X^T}(x^T) - \frac{\alpha}{m}\right)_+}_{P^*_{\zeta^T}(\tilde{\zeta^T})} \right) \in \mathcal{P}(\mathcal{Z}^T),$$

where $P^*_{\zeta^T}(\tilde{\zeta^T}) = \sum_{x^T \in \mathcal{X}^T} \left(P^*_{X^T}(x^T) - \frac{\alpha}{m}\right)_+$.

In particular, if we choose the support as $\mathcal{Z}^T = \mathcal{X}^T \cup \{\tilde{\zeta^T}\}$, the total variation distance between any P_{X^T} and $P^*_{\zeta^T}$ is

$$\mathsf{D_{TV}}(P_{X^T}, P^*_{\zeta^T}) = \sum_{x^T \in \mathcal{X}^T} \left(P_{X^T}(x^T) - \frac{\alpha}{m}\right)_+. \tag{6}$$

In the following, with no risk of confusion, we will refer to $\mathsf{D_{TV}}(P_{X^T}, P^*_{\zeta^T})$ as the quantity defined in (6), even if a different support $\mathcal{Z}^T$ is chosen.

Construct $\mathbb{P}^*_j = P^*_{X^T,\zeta^T|M=j}$ as follows,

$$P^*_{X^T,\zeta^T|M=j}(x^T,\zeta^T) = \begin{cases} P^*_{X^T}(x^T) \wedge P^*_{\zeta^T}(\zeta^T), & \text{if } \gamma^*(x^T,\zeta^T) = j; \\ \frac{\left(P^*_{X^T}(x^T) - P^*_{\zeta^T}(\gamma_j^{*-1}(x^T))\right)_+ \cdot \left(P^*_{\zeta^T}(\zeta^T) - P^*_{X^T}(\gamma_j^{*-1}(\zeta^T))\right)_+}{\mathsf{D_{TV}}(P^*_{X^T}, P^*_{\zeta^T})}, & \text{otherwise,} \end{cases} \tag{7}$$

where γ_j^{*-1} represents the inverse of γ^* for a fixed $j \in [m]$ ($\gamma_j^{*-1}(\zeta^T) = \emptyset$ in particular) and

$$P^*_{X^T} = \operatorname*{arg\,min}_{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T}) \le d} \mathsf{D_{TV}}(P_{X^T}, P^*_{\zeta^T}) = \operatorname*{arg\,min}_{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T}) \le d} \sum_{x^T \in \mathcal{X}^T} \left(P_{X^T}(x^T) - \frac{\alpha}{m}\right)_+.$$

This conditional joint distribution $P^*_{X^T,\zeta^T|M=j}$ with fixed marginals minimizes the j-th error probability $\mathbb{P}_j(\gamma^*(X^T,\zeta^T) \neq j)$, as shown below:

$$\begin{aligned} \mathbb{P}^*_j(\gamma^*(X^T,\zeta^T) \neq j) &= 1 - \sum_{x^T,\zeta^T:\gamma^*(x^T,\zeta^T)=j} (P^*_{X^T}(x^T) \wedge P^*_{\zeta^T}(\zeta^T)) \\ &= 1 - \sum_{x^T,\zeta^T:\gamma^*(x^T,\zeta^T)=j} \left(P^*_{X^T}(x^T) \wedge \frac{\alpha}{m}\right) \\ &= \sum_{x^T \in \mathcal{X}^T} \left(P^*_{X^T}(x^T) - \frac{\alpha}{m}\right)_+, \end{aligned}$$

and ensures that

$$\sum_{\zeta^T} P^*_\zeta(\zeta^T) \mathbb{1}\{\gamma^*(x^T,\zeta^T) \neq 0\} = \sum_{i \in [m]} P^*_\zeta(\zeta^T) \mathbb{1}\{\gamma^*(x^T,\zeta^T) = i\} \le m \cdot \frac{\alpha}{m} = \alpha, \quad \forall x^T.$$

Therefore, the scheme proposed in (7) achieves the min-max j-th error probability in Theorem 4.